DSCI402 Capstone Project: Survey of Time Series

James Craven, Matthew Lindsey, Tina Lane

December 2024

Motivation

1. Diversify conceptual knowledge

Motivation

1. Diversify conceptual knowledge
2. Try different roles in the process

Motivation

1. Diversify conceptual knowledge
2. Try different roles in the process
3. Do something different

Time Series

- **Time Series** - a collection of data indexed by time as opposed to some other values

Time Series

- **Time Series** - a collection of data indexed by time as opposed to some other values
- **Time Series Analyses** - a type of regression that attempts to capture the correlation of the values w.r.t. time.

Project Objectives

1. Learn about the predominant methods for time series analysis

Project Objectives

1. Learn about the predominant methods for time series analysis
2. Analyze the effectiveness of these methods and their pros and cons

Project Objectives

1. Learn about the predominant methods for time series analysis
2. Analyze the effectiveness of these methods and their pros and cons
3. Create materials to demonstrate our learning

The Data

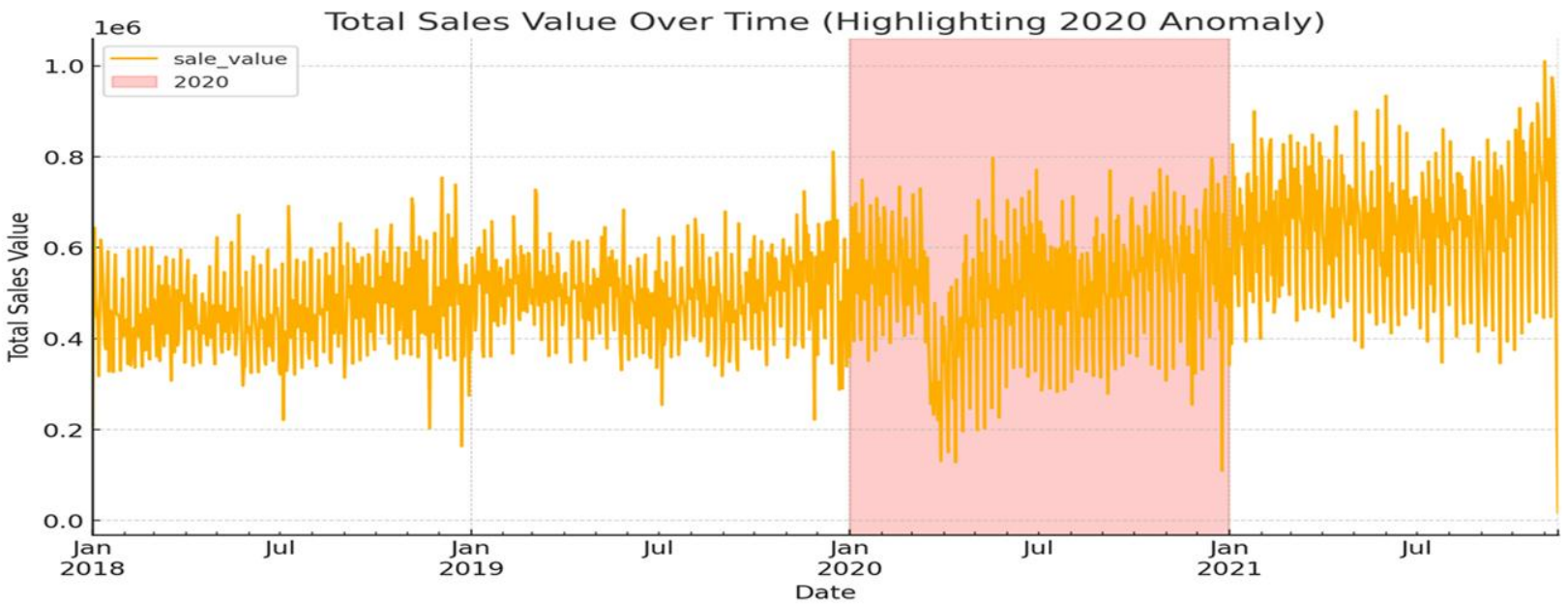
A dataset was provided by a local data science company called Delta Bravo. It is an anonymized, unclean data set that contains a time series representing semi-aggregate motor oil sales data.

The Data

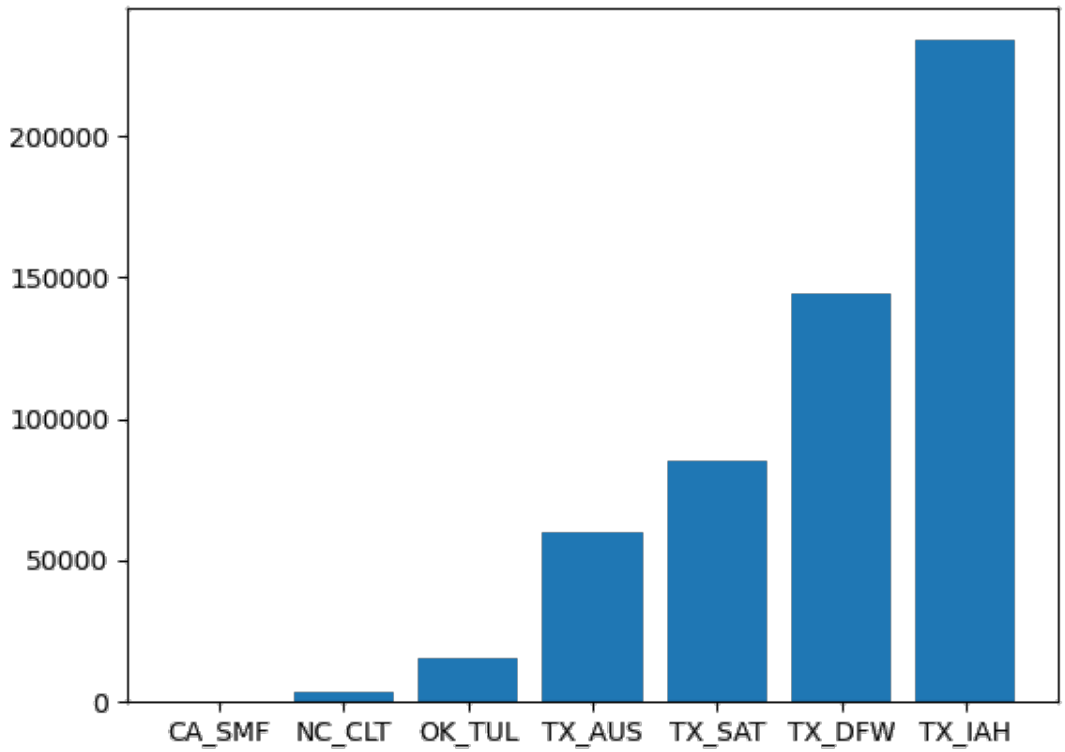
A dataset was provided by a local data science company called Delta Bravo. It is an anonymized, unclean data set that contains a time series representing semi-aggregate motor oil sales data.

1. **Invoice Date** - the date for which the total sales were recorded
2. **Customer Code** - a code that uniquely identifies the customer. These codes were anonymized prior to receiving them in order to protect confidentiality of the original client.
3. **Channel Text** - a label that represented the sales channel. This was almost entirely anonymized, and did not reveal any insights.
4. **Blend** - The type of oil being sold. A categorical attribute that describing whether the oil is conventional or synthetic. This attribute was labelled internally as ”conventional/synthetic”
5. **Variety and Size** - details the oil type and packaging
6. **Sale Value** - the target attribute, the total sales recorded for that row

Data Overview



Spread of Locations



Data Selection

Of the aforementioned only three were preserved

- **Invoice Date** - forms the index of the time series

Data Selection

Of the aforementioned only three were preserved

- **Invoice Date** - forms the index of the time series
- **Sale Value** - forms the values of the time series

Data Selection

Of the aforementioned only three were preserved

- **Invoice Date** - forms the index of the time series
- **Sale Value** - forms the values of the time series
- **Location Code** - allows preservation of regional trends

Data Cleaning

The following steps were taken to handle impurities in the data:

Data Cleaning

The following steps were taken to handle impurities in the data:

- Negative Values - Removed

Data Cleaning

The following steps were taken to handle impurities in the data:

- Negative Values - Removed
- Duplicate Values - Aggregated

Data Cleaning

The following steps were taken to handle impurities in the data:

- Negative Values - Removed
- Duplicate Values - Aggregated
- Missing Values - Filled in and imputed using linear interpolation

Data Cleaning

The following steps were taken to handle impurities in the data:

- Negative Values - Removed
- Duplicate Values - Aggregated
- Missing Values - Filled in and imputed using linear interpolation
- Data set split on location

Methods Used

Two model types were chosen to compare the methods against our data:

Methods Used

Two model types were chosen to compare the methods against our data:

- Statistical approach: **ARIMA**
  - Regression based
  - Explainable

Methods Used

Two model types were chosen to compare the methods against our data:

- Statistical approach: **ARIMA**
  - Regression based
  - Explainable
- Deep Learning approach: **Prophet**
  - Neural Net based
  - Unexplainable - ”black box”

Metrics for evaluation

The data will be split into training and testing sets to be evaluated based on two metrics

Metrics for evaluation

The data will be split into training and testing sets to be evaluated based on two metrics

- Root Mean Square Error (RMSE) - measures predictive accuracy by capturing the average magnitude of error

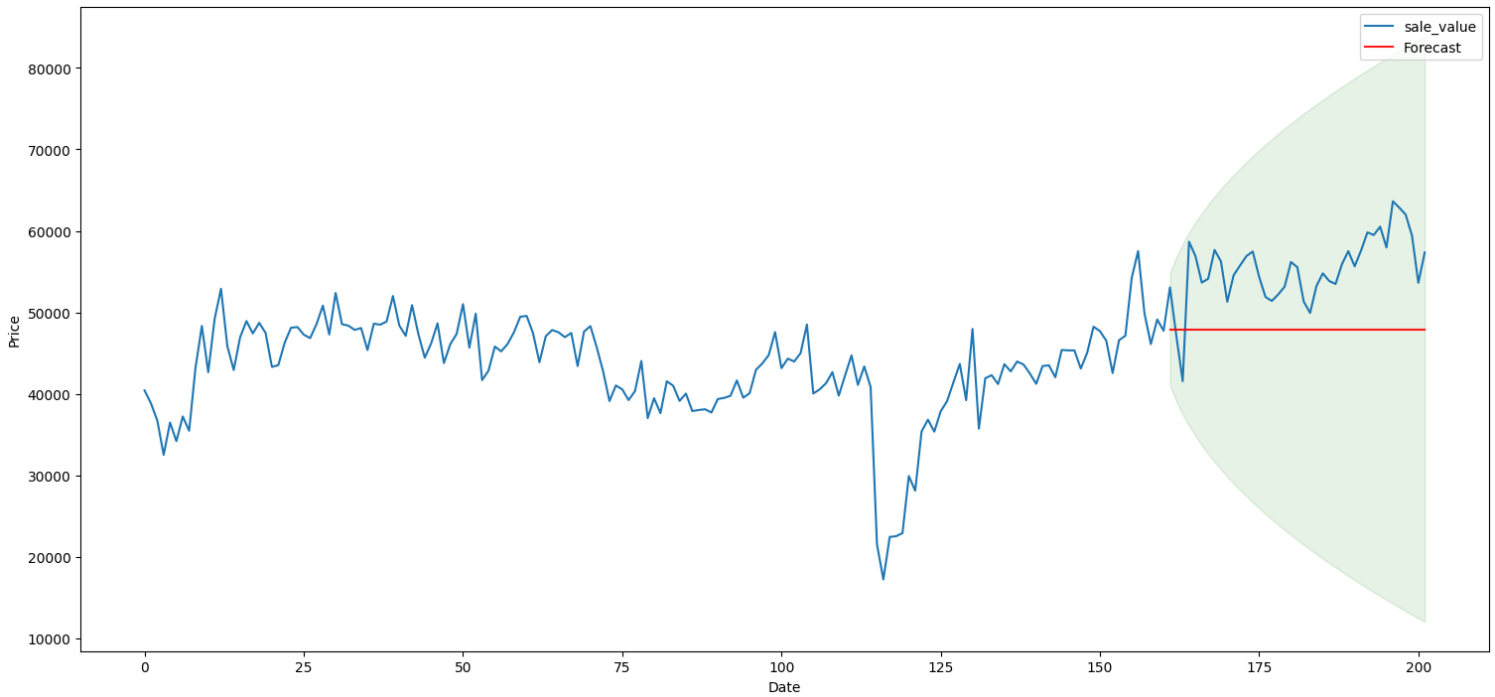
Metrics for evaluation

The data will be split into training and testing sets to be evaluated based on two metrics

- Root Mean Square Error (RMSE) - measures predictive accuracy by capturing the average magnitude of error
- Akaike Information Criterion (AIC) - evaluates the model's goodness of fit while penalizing complexity

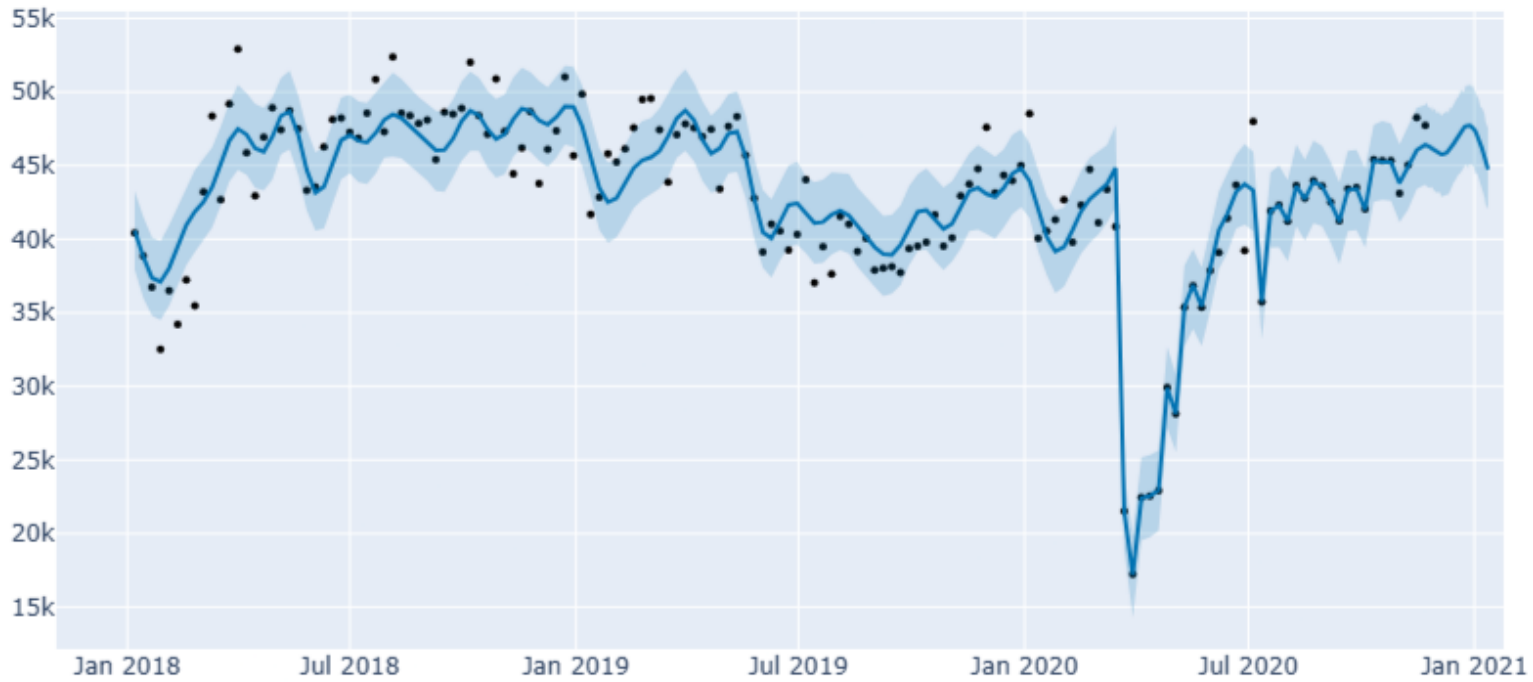
Building ARIMA

AutoARIMA from the Python module ”pmdarima” automates parameter selection based on optimizing AIC scores.



Building Prophet

After training models without accounting for the COVID-19 lockdowns, we tried to incorporate them as custom one-off holiday periods. Those models took into account the dip in sales in 2020.



Model Evaluation

1. Prophet models that incorporated lockdowns outperformed their counterparts.

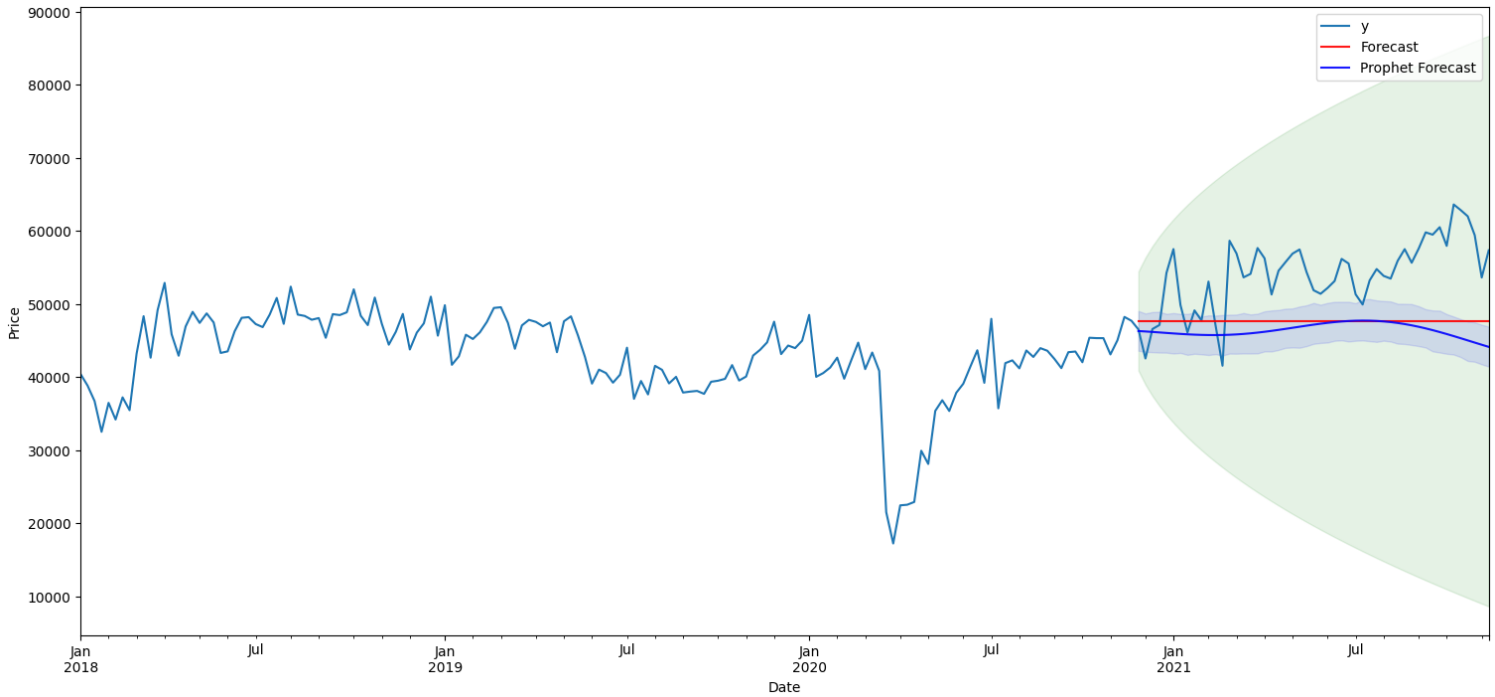
Model Evaluation

1. Prophet models that incorporated lockdowns outperformed their counterparts.
2. ARIMA's linear approach limited its ability to capture non-linear patterns.

Model Evaluation

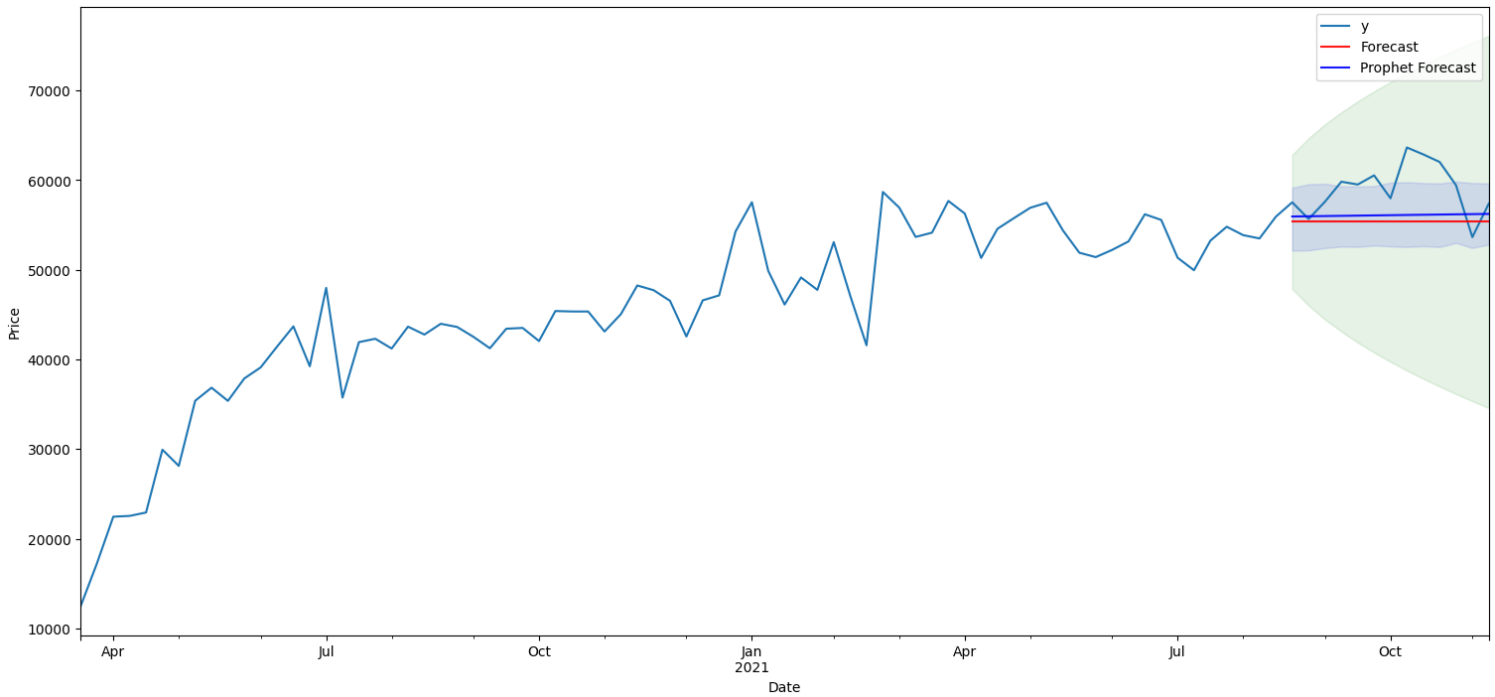
1. Prophet models that incorporated lockdowns outperformed their counterparts.
2. ARIMA's linear approach limited its ability to capture non-linear patterns.
3. Weekly aggregation helped reduce noise, thus had the lowest AIC and RMSE scores.

ARIMA VS Prophet



Finalized Models

The initial test train split was inadequate because the data had a sudden change in trend very close to the split date. To ensure model efficacy the models were also trained on just the original test data to see if it could accurately capture the new trend.



Future Action

- Unused attributes that could be further focused on:
  - Blend
  - Size
  - Variety

Future Action

- Unused attributes that could be further focused on:
  - Blend
  - Size
  - Variety
- Further research into other models used for time-series data

Conclusion

- The project serves as a survey of both ARIMA and Neural Network based models, and as a comparison of the two.

Conclusion

- The project serves as a survey of both ARIMA and Neural Network based models, and as a comparison of the two.
- Both types of models tend to capture similar trends, but with varying levels of granularity and confidence.

Conclusion

- The project serves as a survey of both ARIMA and Neural Network based models, and as a comparison of the two.
- Both types of models tend to capture similar trends, but with varying levels of granularity and confidence.
- ARIMA casts a far wider net, so to speak, and as such values are far more likely to fall in its confidence interval.

Conclusion

- The project serves as a survey of both ARIMA and Neural Network based models, and as a comparison of the two.
- Both types of models tend to capture similar trends, but with varying levels of granularity and confidence.
- ARIMA casts a far wider net, so to speak, and as such values are far more likely to fall in its confidence interval.
- Prophet, on the other hand, tends to be more confident in its prediction and can produce a more nuanced forecast. This allows it to capture smaller, temporary noise slightly better as a result.